Fast and accurate computation of the round-off noise of linear time-invariant systems

J.A. López G. Caffarena C. Carreras O. Nieto-Taladriz

Departamento de Ingeniería Electrónica, ETSI Telecomunicación, Universidad Politécnica de Madrid, Ciudad Universitaria s/n, 28040 Madrid, Spain
E-mail: juanant@die.upm.es

Abstract: From its introduction in the last decade, affine arithmetic (AA) has shown beneficial properties to speed up the time of computation procedures in a wide variety of areas. In the determination of the optimum set of finite word-lengths of the digital signal processing systems, the use of AA has been recently suggested by several authors, but the existing procedures provide pessimistic results. The aim is to present a novel approach to compute the round-off noise (RON) using AA which is both faster and more accurate than the existing techniques and to justify that this type of computation is restricted to linear time-invariant systems. By a novel definition of AA-based models, this is the first methodology that performs interval-based computation of the RON. The provided comparative results show that the proposed technique is faster than the existing numerical ones with an observed speed-up ranging from 1.6 to 20.48, and that the application of discrete noise models leads to results up to five times more accurate than the traditional estimations.

1 Introduction

During the last few years, there has been an increased interest in the determination of the quantisation properties of digital signal processing (DSP) systems using automatic procedures. Among them, the round-off noise (RON) is the most important and most studied finite word-length (FWL) effect of the quantised realisations.

The RON computation procedures are divided into simulation-based, analytical and semi-analytical (or hybrid) ones. The simulation-based procedures are based on performing a very large number of Monte-Carlo simulations, and on computing the statistics of the differences between the quantised signals and their respective unquantised counterparts. They are completely general, but they typically require exceedingly long computation times. In addition, the computed results cannot be reused to compute the FWLs that minimise the RON of the quantised realisation.

The analytical procedures generate an equivalent linear model of the quantised realisation by introducing one uniformly distributed additive white noise (AWN) source per quantiser of the realisation, and propagate the mean and variance of the noise sources through all the signals of the system under study. In most cases, the expressions of the output RON are derived by hand. This approach is well-suited to optimise the specifications of the quantisation operations, but has two major drawbacks. (i) It needs to compute the analytical expressions of the noise gain from each noise source to the output signal, which is a long, tedious and error-prone task; and (ii) as detailed in , the quantised signals must comply with certain assumptions to guarantee the validity of the results. However, in , the authors have suggested an analytical method that automatically computes the expression of the transfer function from each noise source to the output signal, thus performing the required computations in very short time.

The semi-analytical procedures also model the quantisers using AWN sources. However, unlike the analytical case, they perform a reduced number of simulations to compute the parameters of the output

noise. This group of techniques provides the fastest results, but the involved analytical expressions are easily applicable only to linear time-invariant (LTI) systems

Recent research papers have introduced the application of intervals to perform faster computation of the quantisation effects. They perform the computations using affine arithmetic (AA) since it provides tighter bounds of the results than interval arithmetic (IA) However, the published AA-based procedures introduce an exceedingly large number of noise contributions or they are applied to nonlinear systems In both cases, the inclusion of noise symbols more than the ones merely required generates oversized intervals, which cannot be used to accurately compute the statistics of the RON.

The major goal of this paper is to present a new method that performs fast and accurate computation of the RON of LTI systems using AA. This method is based on a semi-analytical interval-based approach. It is the first time that intervals have been used to compute the parameters of the RON, as opposed to computing RON bounds or tails of probability density functions (PDFs), but it will be shown that it provides the fastest results in this type of systems. The second objective of this paper is to justify that the computation of the RON with AA is restricted to LTI systems, since the application of AA to nonlinear systems provides overly pessimistic results. Some additional features of this paper include:

- A new methodology to represent statistical parameters of the noise sources using AA. This methodology has also been extended to model the statistics of more complex signals, such as non-uniform noise models, with AA.

- The introduction of new analytical expressions to accurately characterise the statistical parameters of the rounding noise of previously quantised signals.

- The only reported procedure to compute the RON of LTI systems using AA.

In what follows, Section 2 discusses the related work on the automatic procedures to evaluate the RON. Section 3 introduces the basic concepts of AA. Section 4 explains the proposed AA-based RON computation procedure, laying special emphasis on the generation and propagation of the statistics of noise signals. Section 5 shows the application of the proposed procedure, using two different LTI realisations to detail its correct application. Section 6 compares the accuracy and the speed of the automatic RON evaluation procedures by means of a representative set of examples. Finally, Section 7 presents the conclusions of this work.

2 Related work

Although the analytical evaluation of FWL effects has been studied for a long time, in the last decade a number of research papers have suggested different automatic procedures to speed up the evaluation of the RON In general, these procedures follow a semi-analytical approach, in which: (i) a set of parameterised analytical expressions is developed; and (ii) the values of these parameters are computed using a reduced number of simulations. Hence, the analytical expressions provide the types of required parameters, which depend on the specific procedure, and the simulations provide the values of such elements, which depend on the realisation under study.

According to the expressions used to evaluate the RON, the semi-analytical techniques can be further divided into two groups: (a) techniques that compute the values of the gain factors by means of simulations and (b) techniques that develop series expansions of a given order around the unquantised result.

In the first group, the RON is computed by applying the traditional expressions to evaluate it, but the noise gains associated with the noise sources are computed by performing one simulation per quantiser of the realisation This procedure is very fast and has the advantage of being supported by an extensive theoretical background [7–10]. However, since it makes use of the properties of the LTI systems to minimise the number of simulations, it is difficult to extend its application to nonlinear systems.

In the second group, the most widely used series expansion technique is the Taylor series approximation. This approach has been suggested to perform FWL analysis in , where it has been assumed that the perturbations introduced by the quantisation operations only generate slight deviations of the functionality of the system under study. Several authors have applied first-order approximations of the quantised system using the sensitivity of the output function with respect to the quantisation operations linearisation of the nonlinear operators or second-order approximations Also, the Karhunen-Loéve expansions have recently been suggested to perform a closely related task . In this group of techniques, the authors develop the supporting analytical expressions, but in some cases they recommend performing simpler procedures based on Monte-Carlo simulations to obtain the coefficients of the series expansions These algorithms are theoretically complicated, and significantly slower than the other group techniques, but have been successfully applied to characterise complex nonlinear systems

the authors have also proposed an automatic procedure that provides analytical expressions of the transfer functions from the noise sources to the output

signal. This procedure is based on applying different transformations that successively simplify the graph of the system being evaluated until there is only one element left, whose weight corresponds to the expression of the required transfer function.

In the recent years, several interval-based procedures have been suggested to accelerate the computation of the RON. Among others, multi-IA was one of the approaches originally proposed to estimate the bounds of the quantisation noise This methodology is more accurate than the traditional IA However, it suffers from the same type of overestimation problems, which become particularly important in systems with feedback loops

The most recent interval-based RON evaluation approaches are based on AA. According to their major focus, they are divided into two different groups: (i) they compute the bounds of the behaviour of the quantised system which leads to pessimistic FWLs or (ii) indicate approaches to estimate the tails of the PDF However, none of these approaches compute the statistical parameters of the RON, which is the most commonly used FWL determination procedure

In [15], the authors propose to apply the central limit theorem to estimate the PDF of the output RON, but they introduce one noise source after each arithmetic operation, whose statistical parameters only depend on the quantiser associated with it. However, in some cases (such as sequences of adders of identical FWLs), the intermediate operators do not introduce RON, or it is not accurately described by the traditional noise model The procedure described introduces additional and larger noise terms than those merely required. Thus, the computed parameters of the PDF are overestimated. In this sense, the application of the discrete noise model (DNM) presented later in this paper significantly improves the accuracy of the computations.

the authors indicate that the noise terms of the affine forms are independent among them, and that each one represents the contribution of the source associated with it. However, they apply this procedure to nonlinear systems. A fundamental statement of AA indicates that only the affine operations (addition, subtraction and constant multiplication) are computed exactly, and that it needs to include new noise terms to contain all the possible values of the non-affine operations. This fact implies that the computation of the RON of nonlinear systems (such as polynomials of orders greater than one): (i) leads to an explosion in the number of error symbols, which significantly reduces the speed of the computations and (ii) alters the statistical information associated with the noise terms, generating inaccurate results.

the authors use AA to optimise the target system, guaranteeing at the same time that the new system has identical outputs to the reference one. This procedure yields correct results since the bounds provided by AA always contain all the possible values of the results. However, from a practical point of view, the provided bounds are pessimistic. Instead, it is more realistic to optimise the DSP systems (and particularly the LTI systems) by considering the second-order statistics of the output signals, such as the variance of the output RON

the authors compare the application of different interval-based computation procedures, with emphasis on the accuracy and stability aspects of the AA-based simulations, and suggest a variation of the definition of the quantisation operations of AA to provide tight estimates of the ranges of the signals of quantised LTI systems, particularly when they contain feedback loops. However, this technique is suited to characterise the deterministic evolution of the ranges of the quantised signals, instead of computing their statistical deviations because of the quantisation operations.

3 Affine arithmetic

AA [27] is an extension of the traditional IA [28]. In each operation, AA keeps track of the source and signed amplitude of all the uncertainties that affect each variable. Given affine form $\hat{\boldsymbol{x}}$, its mathematical expression is as follows

$$\hat{\boldsymbol{x}} = x_0 + \sum_{i=1}^{n_x} x_i\, \boldsymbol{\varepsilon}_i, \qquad -1 \leq \boldsymbol{\varepsilon}_i \leq 1 \tag{1}$$

where x_0 is the central value, n_x the number of noise terms and $\boldsymbol{\varepsilon}_i$ and x_i the identifier and amplitude of the ith noise term, respectively In (1), each identifier represents an independent uncertainty contained in the interval $[-1, 1]$. For example, if the input signal is bounded by $[-0.5, 0.5]$, and there is no other information about previous dependencies, its associated affine form is

$$\hat{\boldsymbol{x}} = 0 + 0.5\boldsymbol{\varepsilon}_1 \tag{2}$$

provided that $\boldsymbol{\varepsilon}_1$ has not been previously used.

Given the affine forms $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{b}}$ and constant number c, the description of the basic AA operations is listed in Table 1

Counter $n_{\max}$ represents the value of the largest identifier used by the previous affine forms. In the description of the AA operations shown in Table 1, it is considered that the computations are performed using infinite precision, and hence the effects introduced by the FWL of the machine are considered negligible. Under this assumption, one of the most important features of AA is that the so-called affine operations (addition, subtraction

Table 1 Description of the basic AA operation rules, assuming infinite precision in the computations performed by the machine

Operation	Description of the operation rule
addition	$\hat{a}+\hat{b}=(a_0+b_0)+\sum_{i=1}^{\max(n_a,n_b)}(a_i+b_i)\varepsilon_i$
subtraction	$\hat{a}-\hat{b}=(a_0-b_0)+\sum_{i=1}^{\max(n_a,n_b)}(a_i-b_i)\varepsilon_i$
constant multiplication	$c\,\hat{a}=(c\,a_0)+\sum_{i=1}^{n_a}(c\,a_i)\varepsilon_i$
non-linear multiplication	$\hat{a}\hat{b}=(a_0b_0)+\sum_{i=1}^{\max(n_a,n_b)}(a_0b_i+a_ib_0)\varepsilon_i+\left(\sum_{i=1}^{n_a}\lvert a_i\rvert\sum_{i=1}^{n_b}\lvert b_i\rvert\right)\varepsilon_{n_{\max}+1}$
truncation	$Q_f^{\mathrm{T}}(\hat{a})=(a_0-2^{-f-1})+\left(\sum_{i=1}^{n_a}a_i\varepsilon_i\right)+2^{-f-1}\varepsilon_{n_{\max}+1}$
rounding	$Q_f^{\mathrm{R}}(\hat{a})=a_0+\left(\sum_{i=1}^{n_a}a_i\varepsilon_i\right)+2^{-f-1}\varepsilon_{n_{\max}+1}$

and constant multiplication) are computed exactly, whereas the non-affine operations (i.e. nonlinear operations such as signal multiplication) need to include additional noise terms to contain all the possible values of the results
The quantisation operations (truncation and rounding) are also nonlinear operations. However, it must also be noted that when they are substituted by AWN sources, the quantised descriptions are transformed into linear sequences of operations.

The main advantage of AA is that it alleviates the so-called dependency problem of IA. Consider two input intervals $x=[-1,\ 1]$ and $y=[-1,\ 1]$ that have the relationship $y=-x$. Using IA, the addition of the two intervals yields $z=x+y=[-1,\ 1]+[-1,\ 1]=[-2,\ 2]$, whereas in reality $z=x+y=0$. This simple experiment shows that since IA only manages the bounds of the operands, it is not capable of identifying the cancellations of the sequences of operations.

The computation of these operations using AA is as follows

$$\hat{x}=\varepsilon_1,\quad \hat{y}=-\hat{x}=-\varepsilon_1,\\ \hat{z}=\hat{x}+\hat{y}=\varepsilon_1+(-\varepsilon_1)=0 \tag{3}$$

This result shows that, because of the operation of the noise terms, AA is capable of automatically detecting and cancelling the linear dependencies of the sequences of operations in an efficient manner.

3.1 Example of application

This section describes an example of the application of AA. Let us consider the standard red, green and blue (RGB) to luma and red and blue chroma (YCrCb) converter shown in Fig. 1 , whose sequence of operations is given in Table 2. In it, it is shown how different sets of RGB values are converted by the LTI system. Without loss of generality, in this example, it is considered that the input values are contained in the range [64, 128] in the three signals, and that the computations are performed using infinite precision.

Assuming that the RGB values are independent of each other, the affine forms that represent the signal ranges are modelled using one distinct noise term per uncertainty source, that is

$$\hat{R}=96+32\varepsilon_1,\quad \hat{G}=96+32\varepsilon_2,\quad \hat{B}=96+32\varepsilon_3 \tag{4}$$

By applying the operation rules described in Table 1, the affine form that represents the values of t_1 is

$$\hat{t}_1=0.2220,\quad \hat{R}=0.2220\,(96+32\varepsilon_1)\\ =21.3120+7.1040\varepsilon_1 \tag{5}$$

and the interval that specifies the allowable range of this

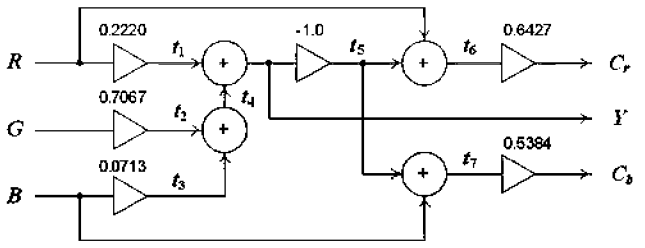

Figure 1 *Detailed description of the standard ITU RGB to YCrCb converter*

Table 2 Description of the computations and signal ranges provided by AA

Operations	Description of the AA-based computations	Signal ranges
$t_1 = 0.2220R$	$\hat{t}_1 = 0.2220(96 + 32\varepsilon_1) = 21.3120 + 7.1040\,\varepsilon_1$	[14.2080, 28.4160]
$t_2 = 0.7067G$	$\hat{t}_2 = 0.7067(96 + 32\varepsilon_2) = 67.8432 + 22.6144\varepsilon_2$	[45.2288, 90.4576]
$t_3 = 0.0713B$	$\hat{t}_3 = 0.0713(96 + 32\varepsilon_3) = 6.8448 + 2.2816\varepsilon_3$	[4.5632, 9.1264]
$t_4 = t_2 + t_3$	$\hat{t}_4 = (67.8432 + 22.6144\varepsilon_2) + (6.8448 + 2.2816\varepsilon_3)$ $= 74.6880 + 22.6144\varepsilon_2 + 2.2816\varepsilon_3$	[49.7920, 99.5840]
$Y = t_1 + t_4$	$\hat{Y} = (21.3120 + 7.1040\varepsilon_1) + (74.6880 + 22.6144\varepsilon_2 + 2.2816\varepsilon_3)$ $= 96 + 7.1040\varepsilon_1 + 22.6144\varepsilon_2 + 2.2816\varepsilon_3$	[64, 128]
$t_5 = -Y$	$\hat{t}_5 = -(96 + 7.1040\varepsilon_1 + 22.6144\varepsilon_2 + 2.2816\varepsilon_3)$ $= -96 - 7.1040\varepsilon_1 - 22.6144\varepsilon_2 - 2.2816\varepsilon_3$	[−128, −64]
$t_6 = t_5 + R$	$\hat{t}_6 = (-96 - 7.1040\varepsilon_1 - 22.6144\varepsilon_2 - 2.2816\varepsilon_3) + (96 + 32\varepsilon_1)$ $= 24.8960\varepsilon_1 - 22.6144\varepsilon_2 - 2.2816\varepsilon_3$	[−49.7920, 49.7920]
$C_r = 0.6427t_6$	$\hat{C}_r = 0.6427 \cdot (24.8960\varepsilon_1 - 22.6144\varepsilon_2 - 2.2816\varepsilon_3)$ $= 16.0007\varepsilon_1 - 14.5343\varepsilon_2 - 1.4664\varepsilon_3$	[−32.0013, 32.0013]
$t_7 = t_5 + B$	$\hat{t}_7 = (-96 - 7.1040\varepsilon_1 - 22.6144\varepsilon_2 - 2.2816\varepsilon_3) + (96 + 32\varepsilon_3)$ $= -7.1040\varepsilon_1 - 22.6144\varepsilon_2 + 29.7184\varepsilon_3$	[−59.4368, 59.4368]
$C_b = 0.5384t_7$	$\hat{C}_b = 0.5384 \cdot (-7.1040\varepsilon_1 - 22.6144\varepsilon_2 + 29.7184\varepsilon_3)$ $= -3.8248\varepsilon_1 - 12.1756\varepsilon_2 + 16.0004\varepsilon_3$	[−32.0008, 32.0008]

affine form is

$$t_1 = t_{1,0} \pm \sum_{i=1}^{n_{t1}} |t_{1,i}| = 21.3127 \pm 7.1040$$
$$= [14.2080, 28.4160] \quad (6)$$

By repeating the same procedure with the rest of the signals of the sequence, the results of the application of AA are given in Table 2, and the ranges associated with the computed affine forms are given in Table 2. This example illustrates that AA is a fast and accurate methodology to evaluate and propagate the signal ranges in LTI systems.

4 Proposed computation procedure

In the proposed AA-based procedure, the RON is computed by propagating the parameters of the PDFs of the noise sources through the linear model of the quantised LTI system. Because of the superposition theorem, the output RON is decomposed into independent contributions, each one described by the mean and variance of its associated source. For this reason, the proposed procedure is divided into three steps: (i) generation of the affine-based models, (ii) propagation of the contributions and (iii) computation of the output statistics.

A general description of this procedure is given in Fig. 2. First, each quantiser of the realisation is modelled by an independent affine form whose central value and noise terms are, respectively, computed as a function of the mean and variance of the quantisation noise of this quantiser. Next, one AA-based simulation is performed to jointly propagate the mean and variance of all the noise sources through all the signals of the realisation. Finally, the information about the contributions of each source is extracted from the noise terms of the output affine form, and the average power of noise is computed.

The remaining parts of this section explain each one of the following three steps: Section 4.1 describes the definition of the parameters of the noise sources using AA; Section 4.2 reveals the conditions required to perform correct propagation of the affine forms and Section 4.3 introduces the expressions that must be applied to calculate the output statistics.

4.1 Definition of the parameters of the noise sources

In the proposed RON computation procedure, each quantiser of the realisation under analysis is associated with a distinct noise term, previously unused, whose amplitude represents the range of the uniformly-distributed PDF of the quantiser associated with it. For example, the affine

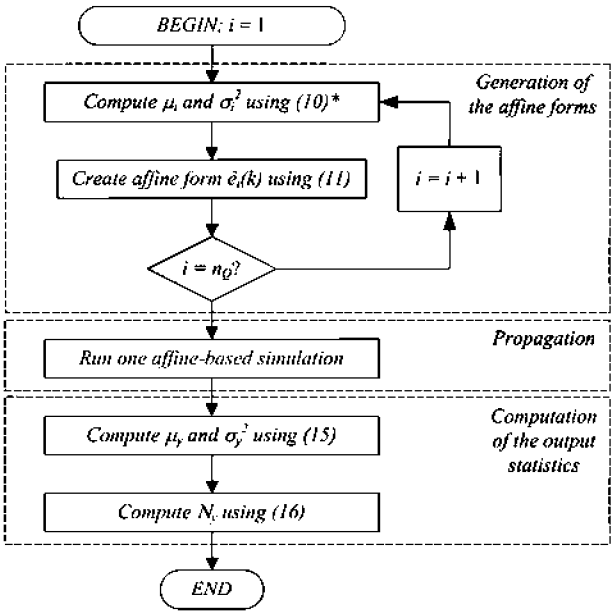

Figure 2 *Proposed affine-based RON computation procedure*

*The use of DNMs to improve the accuracy of the results requires the application of (12) for truncation quantisers and (13) for rounding quantisers instead of (10)

form of (2) indicates that the PDF of x is uniformly distributed in $[-0.5, 0.5]$. Thus, its mean and variance are, respectively, given by

$$m_x = 0, \quad \sigma_x^2 = \frac{(0.5-(-0.5))^2}{12} = \frac{1}{12} \tag{7}$$

Like the other semi-analytical techniques, the proposed procedure applies the superposition theorem to perform independent computation of all the contributions to the output RON. In addition, it also applies the properties of the LTI systems to represent the noise sources using only one noise term per quantiser of the realisation. In LTI systems, the propagation of a given value through the realisation under analysis is time-independent, and so the effects of the propagation of the noise sources need to be computed only once. For this reason, the behaviour of each quantiser is modelled using only one noise term and, consequently, its associated identifier uniquely determines the effects of that source.

For example, consider the quantiser $w_Q = Q_0^{\mathrm{R}}(w)$, which rounds signal w to zero fractional bits. The operation of this quantiser is modelled by adding a given AWN source x to the unquantised signal (i.e. $w_Q = w + x$). In this case, the affine form that represents the distribution of x is as indicated by (2), and the values of the mean and variance of the distribution are given by (7). Note that these values correspond to the sampled time $k = k_0$, but since the statistics of the quantisers do not vary with time, the mean and variance of this source at $k \neq k_0$ are also given by (7).

For notation purposes, the noise signal associated with the quantisation of w is labelled $\boldsymbol{e}_w$. Thus, the operation of the aforementioned quantiser, $w_Q = Q_0^{\mathrm{R}}(w)$, is modelled as $w_Q = w + \boldsymbol{e}_w$. In addition, taking into account the properties of the LTI systems, the noise source $\boldsymbol{e}_w$ is initialised as follows

$$\hat{\boldsymbol{e}}_w(k) = \begin{cases} \hat{\boldsymbol{e}}_w(0), & k = 0 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

where, since the statistics of the distributions are also time-invariant, $\hat{\boldsymbol{e}}_w(0) = \hat{\boldsymbol{e}}_w$.

4.1.1 Use of continuous noise models: Traditionally, the distribution of the values of the AWN sources has been modelled using uniform PDFs Since in this type of PDFs the values of $\boldsymbol{e}_w$ are continuously distributed over its allowable range, we will refer to this case as the continuous noise model (CNM) of the quantisation noise.

Consider that signal w is quantised to f_{w_Q} fractional bits. The general expression of the affine form that represents this CNM is as follows

$$\hat{\boldsymbol{e}}_w = -\frac{2^{-f_{w_Q}}}{2}\xi + \frac{2^{-f_{w_Q}}}{2}\boldsymbol{\varepsilon}_w \tag{9}$$

where ξ is equal to 1 for truncation and 0 for rounding quantisers. It is interesting to note that the mean and variance of this PDF are, respectively, given by

$$m_{\boldsymbol{e}_w} = -2^{-f_{w_Q}-1}\xi, \quad \sigma_{\boldsymbol{e}_w}^2 = \frac{2^{-2f_{w_Q}-2}}{3} \tag{10}$$

where (9) can be rewritten as a function of these parameters as

$$\hat{\boldsymbol{e}}_w = m_{\boldsymbol{e}_w} + \sqrt{3\sigma_{\boldsymbol{e}_w}^2}\,\boldsymbol{\varepsilon}_w \tag{11}$$

Consequently, in the cases where only the mean and variance of the distributions are required, the application of (11) allows the storage of the statistical parameters of the distributions, even of non-uniform PDFs.

4.1.2 Use of DNMs: In the most recent works on the evaluation of the RON, it has been indicated that the models based on uniform PDFs are only strictly valid if the quantisation operations are applied to previously unquantised signals In addition, it has been shown that in fixed-point realisations, the difference between the estimated average power of RON using Monte-Carlo simulations and the traditional CNMs can be up to a 40% For this reason, a new noise model that takes into account the discrete probabilities has been proposed. This model has been called the DNM of the quantisation error However, the expressions of the parameters of the RON have only been derived for

truncation quantisers

$$m_{e_w} = -\frac{1}{2}\left(2^{-f_{w_Q}} - 2^{-f_w}\right), \quad \sigma^2_{e_w} = \frac{1}{12}\left(2^{-2f_{w_Q}} - 2^{-2f_w}\right) \quad (12)$$

where f_w is the number of fractional bits required to represent the signal w without loss of precision.

We have extended this model to provide the corresponding expressions for rounding quantisers

$$m_{e_w} = 0, \quad \sigma^2_{e_w} = \frac{1}{12}\left(2^{-2f_{w_Q}} + 2^{-2f_w+1}\right) \quad (13)$$

provided that $f_w > f_{w_Q}$, and $\sigma^2_{e_w} = 0$, otherwise. The details about the generation of these expressions are provided in the appendix.

It must be noted that the DNMs include the continuous case (in which w is represented using an infinite number of bits), but they are more accurate than the traditional CNMs . This is particularly evident in two situations: (i) in the quantisation of previously quantised signals and (ii) in the multiplication by power-of-two coefficients in which both, the input and the output of the multiplier, are quantised to the same number of fractional bits.

4.2 Propagation of the affine forms

The propagation of the affine forms is performed as indicated by the affine and quantisation operations shown in Table 1. Since AA separately computes the noise terms with different identifiers, the propagation of all the contributions is also performed separately.

However, in each signal, the central value of the computed affine form groups the mean values of all the contributions. Since the contributions of the mean of the noise sources cannot be distinguished from the contributions of the mean of the input signals, this AA-based procedure requires that the mean of the input signals do not affect the computed affine forms. For this reason, if the realisation contains truncation quantisers, all the input signals must be set to zero. If this condition is accomplished, the central value of the computed affine form represents the mean of the RON, and the sum of all the noise terms indicates the shape of its PDF. Note that if the realisation under analysis only contains rounding quantisers, since these quantisers only introduce a new error symbol without affecting the mean value, it is possible to perform a given numerical simulation in combination with the evaluation of the power of the RON.

4.3 Computation of the output statistics

Upon completion of the AA-based simulation, the sequence of affine forms of the output signal y is available. Its expression is as follows

$$\hat{\boldsymbol{y}}(\boldsymbol{k}) = y_0(k) + \sum_{i=1}^{n_y} y_i(k)\boldsymbol{\varepsilon}_i \quad (14)$$

where $y_0(k)$ represents the combined contribution of the means of the noise sources, for each sampled time, to the mean of the output RON at sampled time k; n_y is the number of noise terms of $\hat{\boldsymbol{y}}(\boldsymbol{k})$; $\boldsymbol{\varepsilon}_i$ indicates that the ith noise term is due to the ith noise source and $y_i(k)$ reveals the amount of contribution of that source to the output signal at time k. The mean and variance of the distributions are, respectively, given by

$$m_y(k) = y_0(k), \quad \sigma^2_y(k) = \frac{1}{12}\sum_{i=1}^{n_y}(2y_i(k))^2 = \frac{1}{3}\sum_{i=1}^{n_y} y_i^2(k) \quad (15)$$

and, consequently, the average power of the RON is given by

$$N_y = \sum_{k=-\infty}^{\infty} y_0^2(k) + \frac{1}{3}\sum_{i=1}^{n_y}\sum_{k=-\infty}^{\infty} y_i^2(k) \quad (16)$$

According to the central limit theorem , if the number of contributions exceeds a small number, and the greater values are of the same order of magnitude, the normal PDF is a good approximation to the output RON. Thus, the mean and variance given by (15) and (16) specify this function.

It is important to note that, since in AA the quantisation operations are computed by introducing a new uncertainty source per quantiser of the realisation, the computations performed by the AA-based simulation are identical to the ones indicated by the equivalent linear model of the realisation. For this reason, the proposed approach can be applied to compute the RON of any LTI system.

5 Application examples

5.1 LTI system without feedback loops

This section shows the evaluation of the RON in LTI systems without feedback loops by means of the RGB to YCrCb converter shown in Fig. 1 Without the loss of generality, consider that all the signals of the converter are quantised to a sufficiently large number of integer bits and truncated to a zero fractional bits. The associated sequence of operations is shown in Table 3.

Note that only the multiplications by the constant values generate intermediate results with fractional bits, and hence the quantisation operations only affect the results of these multiplications. In addition, in the five cases $f_w \gg f_{w_Q}$, and, thus, the values of the mean and variance provided by the DNMs are almost equal to the ones obtained using the CNMs. For example, to obtain an accurate representation

Table 3 Description of the computations performed by the proposed procedure and estimated variance of the RON for the RGB to YCrCb converter shown in Fig. 1

Operations	Affine forms computed by the proposed procedure	N_y
$R_Q = Q_0^T(R)$	$\hat{R}_Q = -0.5 + 0.5\varepsilon_1$	0.3333
$G_Q = Q_0^T(G)$	$\hat{G}_Q = -0.5 + 0.5\varepsilon_2$	0.3333
$B_Q = Q_0^T(B)$	$\hat{B}_Q = -0.5 + 0.5\varepsilon_3$	0.3333
$t_1 = Q_0^T(0.2220\ R_Q)$	$\hat{t}_1 = -0.6110 + 0.1110\varepsilon_1 + 0.5\varepsilon_4$	0.4608
$t_2 = Q_0^T(0.7067\ G_Q)$	$\hat{t}_2 = -0.8534 + 0.3534\varepsilon_2 + 0.5\varepsilon_5$	0.8532
$t_3 = Q_0^T(0.0713\ B_Q)$	$\hat{t}_3 = -0.5356 + 0.0356\varepsilon_3 + 0.5\varepsilon_6$	0.3708
$t_4 = t_2 + t_3$	$\hat{t}_4 = -1.3890 + 0.3534\varepsilon_2 + 0.0356\varepsilon_3 + 0.5\varepsilon_5 + 0.5\varepsilon_6$	2.1380
$Y = t_1 + t_4$	$\hat{Y} = -2 + 0.1110\varepsilon_1 + 0.3534\varepsilon_2 + 0.0356\varepsilon_3 + 0.5\varepsilon_4 + 0.5\varepsilon_5 + 0.5\varepsilon_6$	4.2961
$t_5 = -Y$	$\hat{t}_5 = 2 - 0.1110\varepsilon_1 - 0.3534\varepsilon_2 - 0.0356\varepsilon_3 - 0.5\varepsilon_4 - 0.5\varepsilon_5 - 0.5\varepsilon_6$	4.2961
$t_6 = t_5 + R_Q$	$\hat{t}_6 = 1.5 + 0.3890\varepsilon_1 - 0.3534\varepsilon_2 - 0.0356\varepsilon_3 - 0.5\varepsilon_4 - 0.5\varepsilon_5 - 0.5\varepsilon_6$	2.5925
$C_r = Q_0^T(0.6427\ t_6)$	$\hat{C}_r = 0.4641 + 0.25\varepsilon_1 - 0.2271\varepsilon_2 - 0.0229\varepsilon_3 - 0.3214\varepsilon_4 - 0.3214\varepsilon_5 - 0.3214\varepsilon_6 + 0.5\varepsilon_7$	0.4401
$t_7 = t_5 + B_Q$	$\hat{t}_7 = 1.5 - 0.1110\varepsilon_1 - 0.3534\varepsilon_2 + 0.4644\varepsilon_3 - 0.5\varepsilon_4 - 0.5\varepsilon_5 - 0.5\varepsilon_6$	2.6176
$C_b = Q_0^T(0.5384\ t_7)$	$\hat{C}_b = 0.3076 - 0.0598\varepsilon_1 - 0.1902\varepsilon_2 + 0.25\varepsilon_3 - 0.2692\varepsilon_4 - 0.2692\varepsilon_5 - 0.2692\varepsilon_6 + 0.5\varepsilon_8$	0.2845

of 0.7067, at least 13 fractional bits are required. This means that signal t_2 also requires at least 13 fractional bits to be represented without loss of precision (i.e. $f_w = 13$ bits). Since all the signals are truncated to 0 fractional bits ($f_{w_Q} = 0$ bits), the mean and variance of the quantisation error, assuming DNMs, respectively are

$$m_{e_w} = -\frac{1}{2}\left(2^{-f_{w_Q}} - 2^{-f_w}\right) = -\frac{1}{2}(1 - 2^{-13}) \simeq -\frac{1}{2}$$
$$\sigma^2_{e_w} = \frac{1}{12}\left(2^{-2f_{w_Q}} - 2^{-2f_w}\right) = \frac{1}{12}\left(1 - 2^{-26}\right) \simeq \frac{1}{12} \quad (17)$$

which are almost equal to the ones obtained using CNMs ($m_{e_w} = -1/2$, $\sigma^2_{e_w} = 1/12$).

The evaluation of the RON has been performed using the Abaco framework In the set-up process, the simulation tool automatically reads the sequence of operations and the user-defined FWLs, assigns the FWLs of the intermediate signals according to the types of the operations performed and calculates the amount of quantisation performed by each quantiser. Afterwards, the AA-based simulation is executed. Finally, the statistics of the output noise are collected by applying (15) and (16) to the specified signals. The propagation of the affine forms through the sequence of operations and the values of the computed average powers of RON are provided in Table 3.

Consider, for example, signal t_1. Using the affine forms shown in Table 3, the mean, variance and average power of the RON are computed as follows

$$m_{t_1} = -0.6110$$
$$\sigma^2_{t_1} = \frac{1}{3}\sum_{i=1}^{n_{t_1}} t_{1,i}^2 = \frac{1}{3}(0.1110^2 + 0.5^2) = 0.0874 \quad (18)$$
$$N_{t_1} = (m_{t_1})^2 + \sigma^2_{t_1} = 0.4608$$

Since the RON at signal t_1 is composed of two independent noise contributions, the affine form associated with this signal contains only two noise terms. For comparison purposes, Fig. 3 shows the estimated shape of the PDF of the RON and the result of a Monte-Carlo simulation of 10 000 input samples.

The computation of the RON in systems without feedback loops is much simpler and faster than in systems containing feedback loops because the expressions of the mean and variance of (15) do not depend on k, and hence the infinite sums of (16) are removed from this expression. Consequently, in this type of systems, the proposed procedure executes only one iteration of the sequence of operations to provide the results.

5.2 LTI system with feedback loops

Fig. 4 shows the realisation of the second-order infinite impulse response (IIR) filter . This is the first work that has proposed the application of AA to characterise the effects of the FWL on the DSP systems, and hence it will be used here as the second benchmark. Coefficients a_1 and a_2 are, respectively, equal to $-1/\sqrt{2}$

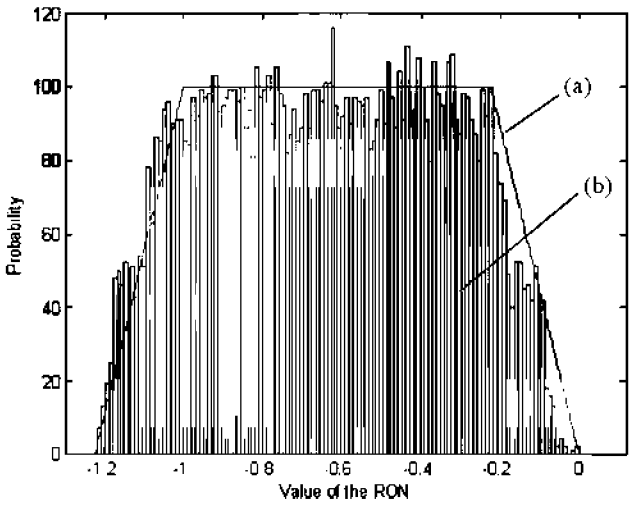

Figure 3 *PDF of the RON of signal t_1 of the RGB to YCrCb converter*

a Theoretical result
b Result of the Monte-Carlo simulation of 10 000 samples

and 1/2, and all the signals of the realisation are quantised to 16 fractional bits using rounding as the underflow strategy. To evaluate the RON, the input signal $x(k)$ is set to zero for all k, and the state variables SV_1 and SV_2 are also initialised to zero.

Note that, similar to the first example, only the input signal and the outputs of the multipliers need quantisers to model the rounding operations to 16 fractional bits. According to the values of the coefficients, the minimum number of fractional bits required to represent the exact values of the signals before and after the quantisers are

$$f_x = \infty, \quad f_{t_1} = 32, \quad f_{t_3} = 17, \quad f_{t_0} = f_{t_2} = f_{t_4} = 16 \quad (19)$$

To develop the equivalent linear model of the realisation, quantisers $Q_1 - Q_3$ are substituted by their associated noise sources $e_1 - e_3$. Note that since the input of Q_3 is only one bit larger than its output, the application of the DNMs provides more accurate results than the traditional CNMs in this case. By applying (13), the statistical parameters of

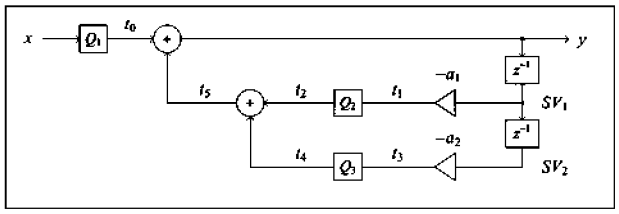

Figure 4 *Detailed description of the IIR realisation evaluated in Section 4*

the noise sources are

$$m_{e_1} = m_{e_2} = m_{e_3} = 0$$
$$\sigma^2_{e_1} = \frac{2^{-32}}{12}, \quad \sigma^2_{e_2} = \frac{1}{12}\left(2^{-32} + 2^{-63}\right) \simeq \frac{2^{-32}}{12} \quad (20)$$
$$\sigma^2_{e_3} = \frac{1}{12}\left(2^{-32} + 2^{-33}\right) = 2.9104 \times 10^{-11}$$

Using (11), these parameters are incorporated in their associated affine forms as follows

$$\hat{e}_i(k)_{(i=1,2,3)} = \begin{cases} 2^{-17}\varepsilon_1, \ 2^{-17}\varepsilon_2, & k = 0 \\ 9.3435 \times 10^{-6}\varepsilon_3, & \\ 0, & \text{otherwise} \end{cases} \quad (21)$$

Table 4 provides the computation of the affine forms at $k = 0$ and 1, the amount of contribution to the average power of the RON at sampled time k, $N_{i,k}$, and the estimated value, N_i, for each case. This example illustrates the propagation of the AA-based models through all the signals of the realisation, as well as the evaluation of the associated contributions of the RON. The successive application of the AA operation rules for $k > 1$ provides the values of the affine forms for the remaining samples and signals of the realisation. Finally, the application of (16) at the end of the simulation provides the value of the average power of the RON of $y(k)$.

Fig. 5 shows the evolution of the estimated values of the output RON as a function of the length of the simulation. As expected from (16), the values provided by the proposed procedure are continuously approaching the exact value. In this case, it can be seen that using only five samples of the simulation the computed result is within the 5% of the final value. However, it must also be noted that for a given degree of accuracy, the number of required samples of the simulation depends on the amount of significant values of $h(k)$ and, consequently, on the proximity of the poles of the filter to $|z| = 1$. Since in this example the number of significant samples of $h(k)$ is small [actually, for $k \geq 15$, $h(k) < h(0)/100$], even a short simulation provides the results with several digits of accuracy. Table 5 provides the computed sequence of output affine forms $\hat{y}(k)$, the contributions to the average power of the RON for each k and the estimated values of the average power of the RON for the initial iteration of this filter.

Fig. 6 also shows the shape of the PDF of the output RON and the result of a Monte-Carlo simulation of 10 000 input samples. In this case, because of the operation of the feedback loops, the RON is composed of an infinite number of independent contributions of the noise sources and, thus, its distribution follows a normal PDF, defined by the average power of the RON given in Fig. 5.

Table 4 Computation of the affine forms at $k = 0$ and 1, the contributions to the average power of RON and the estimated values of the RON for the IIR realisation shown in Fig. 4

Operations	Affine forms computed by the proposed procedure	$N_{i,k}$	N_i
$t_0 = x + e_1$	$\hat{t}_0(0) = 7.63 \times 10^{-6}\varepsilon_1$	1.94×10^{-11}	1.94×10^{-11}
$t_1 = -a_1\,\mathrm{SV}_1$	$\hat{t}_1(0) = 0$	0	0
$t_2 = t_1 + e_2$	$\hat{t}_2(0) = 7.63 \times 10^{-6}\varepsilon_2$	1.94×10^{-11}	1.94×10^{-11}
$t_3 = -a_2\,\mathrm{SV}_2$	$\hat{t}_3(0) = 0$	0	0
$t_4 = t_3 + e_3$	$\hat{t}_4(0) = 9.34 \times 10^{-6}\varepsilon_3$	2.91×10^{-11}	2.91×10^{-11}
$t_5 = t_2 + t_4$	$\hat{t}_5(0) = 7.63 \times 10^{-6}\varepsilon_2 + 9.34 \times 10^{-6}\varepsilon_3$	4.85×10^{-11}	4.85×10^{-11}
$y = t_0 + t_5$	$\hat{y}(0) = 7.63 \times 10^{-6}\varepsilon_1 + 7.63 \times 10^{-6}\varepsilon_2 + 9.34 \times 10^{-6}\varepsilon_3$	6.79×10^{-11}	6.79×10^{-11}
$\mathrm{SV}_2 = z^{-1}\mathrm{SV}_1$	$\widehat{\mathrm{SV}}_2(1) = 0$	0	0
$\mathrm{SV}_1 = z^{-1}y$	$\widehat{\mathrm{SV}}_1(1) = 7.63 \times 10^{-6}\varepsilon_1 + 7.63 \times 10^{-6}\varepsilon_2 + 9.34 \times 10^{-6}\varepsilon_3$	6.79×10^{-11}	6.79×10^{-11}
$t_0 = x + e_1$	$\hat{t}_0(1) = 0$	0	1.94×10^{-11}
$t_1 = -a_1\,\mathrm{SV}_1$	$\hat{t}_1(1) = 5.34 \times 10^{-6}\varepsilon_1 + 5.34 \times 10^{-6}\varepsilon_2 + 6.60 \times 10^{-6}\varepsilon_3$	3.40×10^{-11}	3.40×10^{-11}
$t_2 = t_1 + e_2$	$\hat{t}_2(1) = 5.34 \times 10^{-6}\varepsilon_1 + 5.34 \times 10^{-6}\varepsilon_2 + 6.60 \times 10^{-6}\varepsilon_3$	3.40×10^{-11}	5.34×10^{-11}
$t_3 = -a_2\,\mathrm{SV}_2$	$\hat{t}_3(1) = 0$	0	0
$t_4 = t_3 + e_3$	$\hat{t}_4(1) = 0$	0	2.91×10^{-11}
$t_5 = t_2 + t_4$	$\hat{t}_5(1) = 5.34 \times 10^{-6}\varepsilon_1 + 5.34 \times 10^{-6}\varepsilon_2 + 6.60 \times 10^{-6}\varepsilon_3$	3.40×10^{-11}	8.25×10^{-11}
$y = t_0 + t_5$	$\hat{y}(1) = 5.34 \times 10^{-6}\varepsilon_1 + 5.34 \times 10^{-6}\varepsilon_2 + 6.60 \times 10^{-6}\varepsilon_3$	3.40×10^{-11}	10.19×10^{-10}
$\mathrm{SV}_2 = z^{-1}\mathrm{SV}_1$	$\widehat{\mathrm{SV}}_2(2) = 7.63 \times 10^{-6}\varepsilon_1 + 7.63 \times 10^{-6}\varepsilon_2 + 9.34 \times 10^{-6}\varepsilon_3$	6.79×10^{-11}	6.79×10^{-11}
$\mathrm{SV}_1 = z^{-1}y$	$\widehat{\mathrm{SV}}_1(2) = 5.34 \times 10^{-6}\varepsilon_1 + 5.34 \times 10^{-6}\varepsilon_2 + 6.60 \times 10^{-6}\varepsilon_3$	3.40×10^{-11}	10.19×10^{-10}

6 Comparison between the automatic RON evaluation procedures

This section compares the performance of the proposed procedure with that of the procedures described in the related bibliography, in terms of computational speed and accuracy of the results. Section 6.1 details the selection and preparation of the experiments and Section 6.2 provides a comparative discussion about the results obtained in each case.

6.1 Selection and preparation of the experiments

To assess the performance of the proposed approach, the proposed procedure has been applied to compute the RON of a wide variety of LTI realisations. Among them, some examples published in the related literature have been selected to compare its accuracy and computational speed.

Tables 6 and 7, respectively, show the average power of RON, N_y, and the computation times, in seconds, required by the simulation-based approach the semi-analytical approach and the proposed approach. Table 6 also shows the values of N_y provided by the analytical approach

The computation of the RON has been performed using five LTI realisations of different complexities: (i) the RGB to YCrCb converter (Conv.)

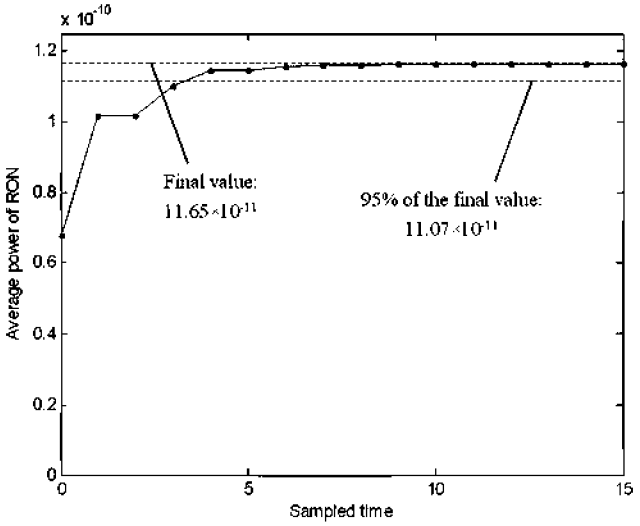

Figure 5 *Accuracy of the estimation of the output RON of the IIR realisation described in Fig. 4 as a function of the length of the simulation*

Table 5 Evolution of the values of the affine forms, the contributions to the average power of RON and the estimated values of the RON for the output signal y of the IIR realisation shown in Fig. 4

k	Computed output affine forms	$N_{y,k}$	N_y
0	$\hat{y}(0) = 7.63 \times 10^{-6}\varepsilon_1 + 7.63 \times 10^{-6}\varepsilon_2 + 9.34 \times 10^{-6}\varepsilon_3$	6.79×10^{-11}	6.79×10^{-11}
1	$\hat{y}(1) = 5.34 \times 10^{-6}\varepsilon_1 + 5.34 \times 10^{-6}\varepsilon_2 + 6.60 \times 10^{-6}\varepsilon_3$	3.40×10^{-11}	10.19×10^{-11}
2	$\hat{y}(2) = 0$	0	10.19×10^{-11}
3	$\hat{y}(3) = -2.70 \times 10^{-6}\varepsilon_1 - 2.70 \times 10^{-6}\varepsilon_2 - 3.30 \times 10^{-6}\varepsilon_3$	8.49×10^{-12}	11.03×10^{-11}
4	$\hat{y}(4) = -1.91 \times 10^{-6}\varepsilon_1 - 1.91 \times 10^{-6}\varepsilon_2 - 2.33 \times 10^{-6}\varepsilon_3$	4.24×10^{-12}	11.46×10^{-11}
5	$\hat{y}(5) = 0$	0	11.46×10^{-11}
6	$\hat{y}(6) = 9.54 \times 10^{-7}\varepsilon_1 + 9.54 \times 10^{-7}\varepsilon_2 + 1.17 \times 10^{-6}\varepsilon_3$	1.06×10^{-12}	11.57×10^{-11}
7	$\hat{y}(7) = 6.74 \times 10^{-7}\varepsilon_1 + 6.74 \times 10^{-7}\varepsilon_2 + 8.26 \times 10^{-7}\varepsilon_3$	5.30×10^{-13}	11.62×10^{-11}
8	$\hat{y}(8) = 0$	0	11.62×10^{-11}
9	$\hat{y}(9) = -3.37 \times 10^{-7}\varepsilon_1 - 3.37 \times 10^{-7}\varepsilon_2 - 4.13 \times 10^{-7}\varepsilon_3$	1.33×10^{-13}	11.63×10^{-11}
10	$\hat{y}(10) = -2.38 \times 10^{-7}\varepsilon_1 - 2.38 \times 10^{-7}\varepsilon_2 - 2.92 \times 10^{-7}\varepsilon_3$	6.63×10^{-14}	11.64×10^{-11}
11	$\hat{y}(11) = 0$	0	11.64×10^{-11}
12	$\hat{y}(12) = 1.19 \times 10^{-7}\varepsilon_1 + 1.19 \times 10^{-7}\varepsilon_2 + 1.46 \times 10^{-7}\varepsilon_3$	1.65×10^{-14}	11.64×10^{-11}
13	$\hat{y}(13) = 8.43 \times 10^{-8}\varepsilon_1 + 8.43 \times 10^{-8}\varepsilon_2 + 1.03 \times 10^{-7}\varepsilon_3$	8.29×10^{-15}	11.64×10^{-11}
14	$\hat{y}(14) = 0$	0	11.64×10^{-11}
15	$\hat{y}(15) = -4.21 \times 10^{-8}\varepsilon_1 - 4.21 \times 10^{-8}\varepsilon_2 - 5.15 \times 10^{-8}\varepsilon_3$	2.07×10^{-15}	11.64×10^{-11}

detailed in (ii) the eight-point inverse discrete cosine transform (IDCT8) (iii) the second-order IIR filter (IIR2) (iv) the third-order basic lattice filter (Lat3) studied in [10, p. 434] and (v) the sixth-order transposed direct form II filter based on the modified delta-operator (ρDFIIt6) with minimum RON Case studies (i) and (ii) are multi-output image processing systems without feedback loops. In these two cases, N_y represents the sum of the average power of RON of all the output signals. Case studies (iii)–(v) are IIR filters of different complexities. For each structure, the type of response (finite impulse response (FIR) or IIR), and the number of operations of the realisation (n. ops.) are also given. In addition, Table 7 also provides the minimum speed-up obtained by the proposed approach with respect to the other approaches in each case.

In the five realisations, all the signals are assumed to have a sufficiently large number of integer bits, and they are all rounded to 16 fractional bits. In the simulation-based method, each provided RON value has been computed by averaging the results of 10 simulations of 10^6 samples long. In all cases, the results of the individual simulations have been identical to their corresponding averaged RON values up to the second significant digit, and in most cases up to the third significant digit. The analytical and semi-analytical methods assume that all the noise sources have uniform PDFs in their respective ranges of existence (i.e. they use the traditional CNMs). In the proposed procedure, the values of the noise sources take into account the previous quantisation operations (i.e. they apply the proposed DNMs). In the semi-analytical and the proposed methods, the length of the simulations has been set to 100 samples in each case to compute the

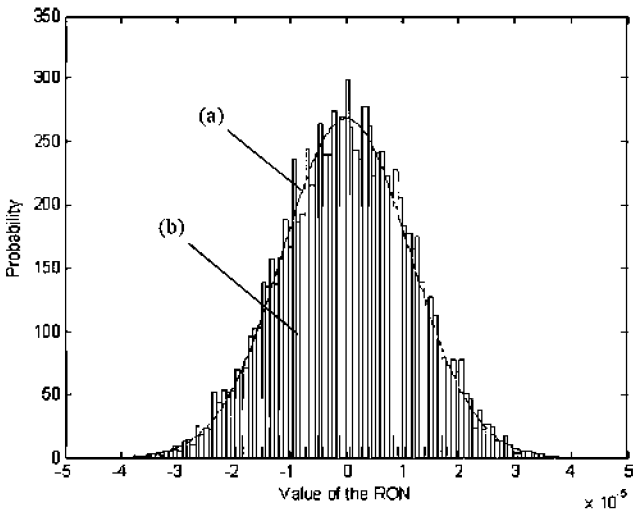

Figure 6 *PDF of the output RON of the IIR realisation*

a Theoretical result

b Result of the Monte-Carlo simulation of 10 000 samples

Table 6 Average power of RON provided by the four types of RON evaluation procedures

LTI system	Type of response	n. ops.	Monte-Carlo simulation	Analytical (with CNMs)	Semi-analytical (with CNMs)	Proposed (with DNMs)
Conv.	FIR	9	1.66×10^{-10}	1.66×10^{-10}	1.66×10^{-10}	1.66×10^{-10}
IDCT8	FIR	36	6.54×10^{-9}	6.54×10^{-9}	6.54×10^{-9}	6.54×10^{-9}
IIR2	IIR	4	1.21×10^{-10}	0.99×10^{-10}	0.99×10^{-10}	1.16×10^{-10}
Lat3	IIR	19	8.54×10^{-11}	8.47×10^{-11}	8.47×10^{-11}	8.47×10^{-11}
ρDFIIt6	IIR	37	7.60×10^{-11}	7.97×10^{-11}	7.97×10^{-11}	7.97×10^{-11}

noise gain from each noise source to the output signal. However, the semi-analytical procedure performs one numerical simulation per AWN source of the realisation and, thus, the simulator is invoked and loaded in memory once per quantiser of the realisation. Instead, the proposed AA-based approach performs only one simulation to compute all the results. The proposed approach has been implemented in our FWL evaluation framework, called Abaco and the computations have been run on an Intel Pentium IV PC processor running at 2.8 GHz with 2 GB of RAM memory and Linux Red Hat 7.3 operating system.

6.2 Discussion

From the results shown in Tables 6 and 7, the following points are derived.

1. The analytical, semi-analytical and proposed methods provide very similar results in most cases. This fact occurs because the three methods substitute the quantisers by independent noise sources, whose values are propagated according to different rules. Therefore they use different techniques to compute the results, but since the statistical parameters of the noise sources are identical, they provide the same results.

2. In some cases, such as the IIR2 filter, the analytical and the semi-analytical procedures provide results slightly different from those of the Monte-Carlo simulation and the proposed approach. This difference appears because the CNMs of the noise sources do not consider the previous quantisation operations. Owing to the application of the DNMs of (13), the error of the estimator is reduced from 21% to 4% with respect to the Monte-Carlo simulation and, thus, the proposed estimator is more than five times more accurate than the existing one in this case. A similar result was obtained for truncation quantisers but the expressions provided here for rounding quantisers complete this work.

3. The simulation-based procedure is slow, but it guarantees to provide accurate results. In fact, if the simulations are long enough, this technique provides the reference values [42]. Compared with this, the proposed method provides fast and precise approximations to the reference results.

4. The proposed procedure is the fastest method. It provides results up to 20 times faster than the semi-analytical method in the systems without feedback loops, and approximately 1.6 times faster than the IIR filters. In both cases, the ratio increases with the number of operations. This fact occurs because the proposed approach performs only one AA-based simulation in all cases, whereas the number of computations of the other approaches increases with the number of noise sources of the realisation under study.

Table 7 Computation times of the four types of RON evaluation procedures, and minimum speed-up obtained by the proposed procedure with respect to the other approaches

LTI system	Type of response	n. ops.	Monte-Carlo simulation	Semi-analytical (with CNMs)	Proposed (with DNMs)	Minimum speed up
Conv.	FIR	9	7.3×10^{2}	6.4×10^{-1}	8.1×10^{-2}	7.90
IDCT8	FIR	36	2.8×10^{3}	4.3×10^{0}	2.1×10^{-1}	20.48
IIR2	IIR	4	3.4×10^{2}	1.6×10^{-1}	1.0×10^{-1}	1.60
Lat3	IIR	19	5.3×10^{2}	7.7×10^{-1}	4.6×10^{-1}	1.67
ρDFIIt6	IIR	37	1.0×10^{3}	3.2×10^{0}	1.9×10^{0}	1.68

The analytical approach of also provides very fast and accurate results. However, the published computation times only refer to a part of the process, and so they are not directly comparable with the ones provided in Table 7. the expressions of the transfer functions associated with the noise sources of the IIR2 filter are obtained in 0.08 s. In the proposed approach, the equivalent part of the process is the execution of the AA-based simulation, which only requires 0.03 s.

However, since the experiments are not performed under identical conditions in both cases, these values should be considered only as an approximation.

7 Conclusions

This contribution has presented a new method to compute the RON of quantised LTI systems using AA. The analysis of the related literature has revealed that the AA-based FWL determination procedures provide estimates of the bounds of the PDFs, but there are no published AA-based RON computation procedures. In this context, a new semi-analytical procedure to compute the RON has been detailed. It has been justified that its correct application is restricted to LTI systems, and it has been combined with novel expressions of the DNM of rounding quantisers, which complement the existing ones of truncation quantisers. The examples have shown that the discrete noise model is up to five times more accurate than the traditional continuous model, and that the AA-based computation procedure faster than other semi-analytic methods, with an observed speed-up of 1.6 in simple IIR filters and up to 20.48 in complex FIR realisations.

8 Acknowledgment

This work was supported in part by the Spanish Ministry of Education and Science, under projects TIC2003-09061-C03-02 and TEC2006-13067-C03-03.

9 References

TODMAN T.J., CONSTANTINIDES G.A., WILTON S.J.E., MENCER O., LUK W., CHEUNG P.Y.K.: 'Reconfigurable computing: architectures and design methods', *IEE Proc., Comput. Digit. Tech.*, 2005, **152**, pp. 193–207

SUNG W., KUM K.-I.: 'Simulation-based word-length optimization method for fixed-point digital signal processing systems', *IEEE Trans. Signal Process.*, 1995, **43**, pp. 3087–3090

WILLEMS M., BURSGENS V., KEDING H., GROTKER T., MEYR H.: 'System level fixed-point design based on an interpolative approach'. Proc. 34th Design Autom. Conf., 1997, pp. 293–298

CMAR R., RIJNDERS L., SCHAUMONT P., VERNALDE S., BOLSENS I.: 'A methodology and design environment for DSP ASIC fixed point refinement'. Proc. Design, Autom. Test Europe Conf. and Exhib, 1999, pp. 271–276

KUM K.-I., SUNG W.: 'Combined word-length optimization and high-level synthesis of digital signal processing systems', *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, 2001, **20**, pp. 921–930

ROY S., BANERJEE P.: 'An algorithm for trading off quantization error with hardware resources for MATLAB-based FPGA design', *IEEE Trans. Comput.*, 2005, **54**, pp. 886–896

JACKSON L.B.: 'Digital filters and signal processing' (Kluwer Academic Publishers, Boston, 1986)

ROBERTS R.A., MULLIS C.T.: 'Digital signal processing' (Addison-Wesley, Reading, MA, 1987)

OPPENHEIM A.V., SCHAFER R.W.: 'Discrete-time signal processing' (Prentice Hall, Englewood Cliffs, NJ, 1989)

PARHI K.K.: 'VLSI digital signal processing systems: design and implementation' (Wiley, New York, 1999)

MENARD D., SENTIEYS O.: 'Automatic evaluation of the accuracy of fixed-point algorithms'. Proc. Design, Autom. Test in Europe Conf. Exhib., 2002, Paris, France, March 2002, pp. 529–535

CONSTANTINIDES G.A., CHEUNG P.Y.K., LUK W.: 'Roundoff-noise shaping in filter design'. Proc. IEEE Int. Symp. Circ. Systems, 2000, vol. 4, pp. 57–60

CONSTANTINIDES G.A.: 'Perturbation analysis for word-length optimization'. Proc. 11th IEEE Symp. Field-Programm. Custom Comput. Mach., 2003

CONSTANTINIDES G.A., CHEUNG P.Y.K., LUK W.: 'Wordlength optimization for linear digital signal processing', *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, 2003, **22**, pp. 1432–1442

FANG C.F., RUTENBAR R.A., CHEN T.: 'Fast, accurate static analysis for fixed-point finite-precision effects in DSP designs'. Proc. Int. Conf. Computer-Aided Design, 2003 (ICCAD '03), 2003

GAFFAR A.A., MENCER O., LUK W.: 'Unifying bit-width optimisation for fixed-point and floating-point designs'. Proc. 12th IEEE Symp. Field-Programm. Custom Comput. Mach. (FCCM 2004), 2004, pp. 79–88

CHAN S.C., TSUI K.M.: 'Wordlength determination algorithms for hardware implementation of linear time invariant systems with prescribed output accuracy'. Proc. IEEE Int. Symp. Circuits Systems, 2005 (ISCAS 2005), 2005

HAN K., EVANS B.L.: 'Optimum wordlength search using sensitivity information', *EURASIP J. Appl. Signal Process.*, 2006, **2006**, (14), article ID 92849, 14 pp.

MENARD D., SENTIEYS O.: 'A methodology for evaluating the precision of fixed-point systems'. Proc. IEEE Int Conf. Acoustics, Speech, Signal Process. (ICASSP '02), 2002, vol. 3, pp. III-3152–III-3155

MENARD D., CHILLET D., SENTIEYS O.: 'Floating-to-fixed-point conversion for digital signal processors', *EURASIP J. Appl. Signal Process.*, 2006, **2006**, (14), article ID 96421, 19 pp.

SHI C., BRODERSEN R.W.: 'A perturbation theory on statistical quantization effects in fixed-point DSP with non-stationary inputs'. Proc. Int. Symp. Circuits Systems 2004 (ISCAS '04), 2004

SHI C., BRODERSEN R.W.: 'Floating-point to fixed-point conversion with decision errors due to quantization'. Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process., 2004 (ICASSP '04), 2004

WU B., ZHU J., NAJM F.N.: 'Dynamic range estimation', *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, 2006, **25**, pp. 1618–1636

LÓPEZ J.A.: 'Evaluación de los Efectos de Cuantificación en las Estructuras de Filtros Digitales Utilizando Técnicas de Cuantificación Basadas en Extensiones de Intervalos', PhD Thesis, Dept. Ingeniería Electrónica, Univ. Politécnica de Madrid, 2004 Madrid

LOPEZ J.A., CARRERAS C., CAFFARENA G., NIETO-TALADRIZ O.: 'Fast characterization of the noise bounds derived from coefficient and signal quantization'. Proc. Int. Symp. Circuits Systems 2003 (ISCAS '03), 2003

LOPEZ J.A., CARRERAS C., NIETO-TALADRIZ O.: 'Improved interval-based characterization of fixed-point LTI systems with feedback loops', *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, 2007, **26**, pp. 1923–1933

STOLFI J., DE FIGUEIREDO L.H.: 'Self-validated numerical methods and applications'. Proc. 21st Brazilian Mathematics Colloquium, IMPA, Rio de Janeiro, Brazil, 1997

MOORE R.E.: 'Interval analysis' (Prentice-Hall, Englewood Cliffs, NJ, 1966)

LEE D.-U., GAFFAR A.A., CHEUNG R.C.C., MENCER O., LUK W., CONSTANTINIDES G.A.: 'Accuracy-guaranteed bit-width optimization', *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, 2006, **25**, pp. 1990–2000

PU Y., HA Y.: 'An automated, efficient and static bit-width optimization methodology towards maximum bit-width-to-error tradeoff with affine arithmetic model'. Proc. Asia and South Pacific Design Autom. Conf. (ASPDAC '06), 2006

WADEKAR S.A., PARKER A.C.: 'Accuracy sensitive word-length selection for algorithm optimization'. Proc. Int. Conf. Computer Design (ICCD '98), 1998

WU B., ZHU J., NAJM F.N.: 'An analytical approach for dynamic range estimation'. Proc. 41st Design Autom. Conf., 2004

SHI C., BRODERSEN R.W.: 'Automated fixed-point data-type optimization tool for signal processing and communication systems'. Proc. 41st Design Autom. Conf., 2004

CARRERAS C., LOPEZ J.A., NIETO-TALADRIZ O.: 'Bit-width selection for data-path implementations'. Proc. 12th Int. Symp. System Synthesis, 1999

BENEDETTI A., PERONA P.: 'Bit-width optimization for configurable DSP's by multi-interval analysis'. Proc. 34th Asilomar Conf. Signals, Systems Comp., 2000

CONSTANTINIDES G.A., CHEUNG P.Y.K., LUK W.: 'Truncation noise in fixed-point SFGs', *Electron. Lett.*, 1999, **35**, pp. 2012–2014

LOPEZ J.A., CAFFARENA G., CARRERAS C.: 'Rounding Noise in fixed-point SFGs'. Technical Report, Univ. Politecnica de Madrid, Madrid, 2007

CARRERAS C., WALKER I.D.: 'Interval methods for fault-tree analysis in robotics', *IEEE Trans. Reliab.*, 2001, **50**, pp. 3–11

WALKER I.D., CARRERAS C., McDONNELL R., GRIMES G.: 'Extension versus bending for continuum robots', *Int. J. Adv. Robot.*, 2006, **3**, pp. 171–178

LÓPEZ J.A., CARRERAS C., NIETO-TALADRIZ O.: 'From Matlab floating-point algorithms to VHDL fixed-point specifications: a concurrent methodology'. Proc. DSP Deutschland '99, 1999

LI G., ZHAO Z.: 'On the generalized DFIIt structure and its state-space realization in digital filter implementation', *IEEE Trans. Circuits Syst. I, Regul. Pap.*, 2004, **51**, pp. 769–778

JERUCHIM M.: 'Techniques for estimating the bit error rate in the simulation of digital communication systems', *IEEE J. Sel. Areas Commun.*, 1984, **2**, pp. 153–170

10 Appendix: Derivation of the expressions of the mean and variance for rounding quantisers

Figs. 7*a*–7*i* show the PDFs of the quantisation of previously quantised signals under different amounts of quantisation specifications They have been obtained by connecting in series two quantisers, namely Q_1 and Q_2, as described below

$$w = Q_{f_w}(x), \quad w_Q = Q_0^R(w) \tag{22}$$

The input variable x follows a uniform distribution in $[-10^3, 10^3]$, and the computations have been performed in MATLAB. Quantiser Q_1 selects the number of fractional bits of the input signal, and Q_2 always performs rounding to zero fractional bits. Therefore the output quantisation error is bounded by $[-0.5, 0.5]$ and its mean is equal to 0 in all cases (i.e. $m_{e_w} = 0$). In order to derive the expression of the variance for rounding quantisers, the following characteristics must be considered *a priori*.

1. The type of the PDF only depends on the difference $\Delta f = f_w - f_{w_Q}$. In other words, the shape of the PDF is maintained regardless of the number of bits of Q_2. Consequently, for a given Δf, if Q_2 performs the quantisation operations to f_{w_Q} fractional bits instead of 0, the only effect on the result is to multiply the computed value by $0.25 f_{w_Q}$. Hence, the expression of the output variance follows the relationship

$$\sigma^2_{f_{w_Q},\Delta f} = \sigma^2_{0,\Delta f}\, 2^{-2f_{w_Q}} \tag{23}$$

where $\sigma^2_{f_{w_Q},\Delta f}$ and $\sigma^2_{0,\Delta f}$, respectively, represent the variances of the DNMs considering that Q_2 performs the quantisation operations to f_{w_Q} or to 0 fractional bits. For this reason, the number of fractional bits of Q_2 in (22) has been fixed to 0.

2. By applying the definition of the variance, the numerical results of $\sigma^2_{0,\Delta f}$ for the initial values of $i = \Delta f$ are given in Table 8. Once these values have been obtained, it is important to note that they require an increasing number of fractional bits for its exact representation. For example, 0.125 only requires three fractional bits, 0.09375 requires five fractional bits and so on. The computation of the equivalent fractions of these values reveals that the denominators follow a simple progression of the form 2^{2i+1}. In addition, a further insight reveals that the numerators also follow a geometric series of the form

$$\begin{aligned} \mathrm{Num}_i &= 1 + 2 + 2^3 + 2^5 + 2^7 + \cdots + 2^{2i-1} \\ &= 1 + 2\sum_{j=0}^{i-2} 2^{2j}, \quad i \geq 1 \end{aligned} \tag{24}$$

where Num_i is the numerator of the ith fraction of the sequence. Consequently, the general expression to obtain

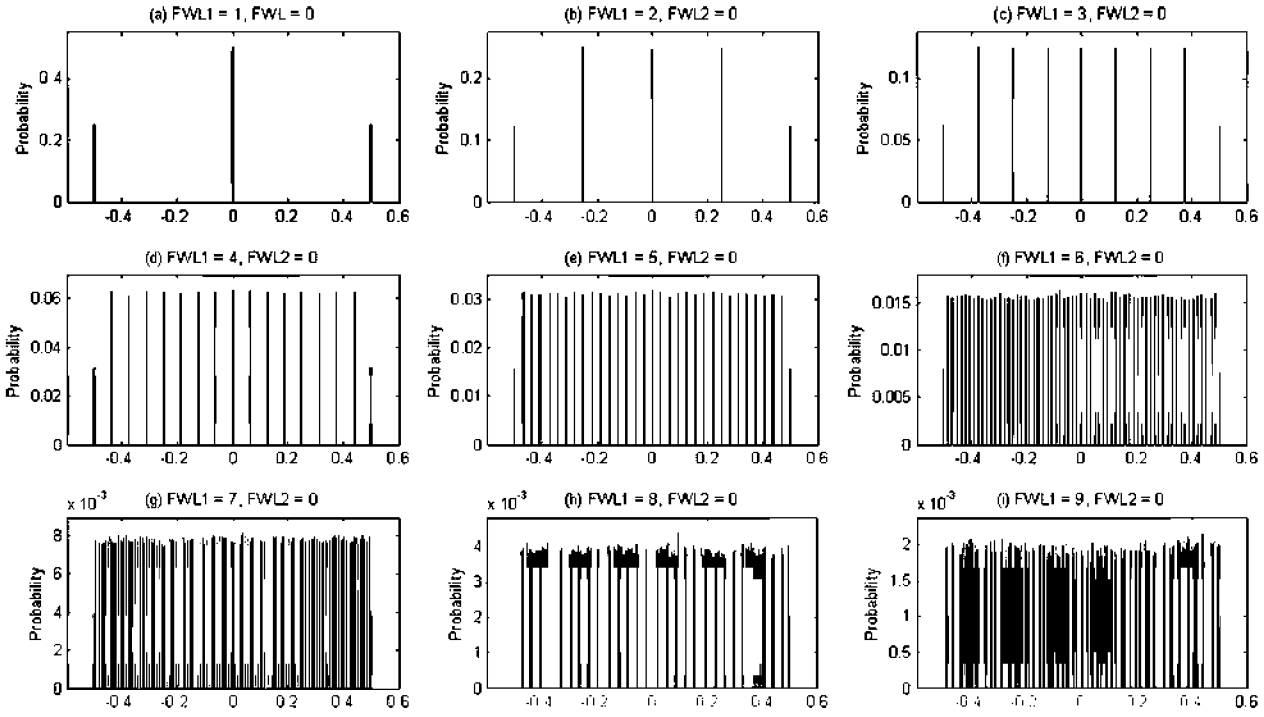

Figure 7 *Discrete PDFs of the quantisation of previously quantised signals, as a function of the FWLs assigned by the two consecutive quantisers*

Table 8 Values of the variance of the PDFs described in Fig. 7

Δf	$\sigma^2_{0,\Delta f}$	Equivalent fraction
1	0.12500000000000	$1/2^3$
2	0.09375000000000	$3/2^5$
3	0.08593750000000	$11/2^7$
4	0.08398437500000	$43/2^9$
5	0.08349609375000	$171/2^{11}$
6	0.08337402343750	$683/2^{13}$
7	0.08334350585938	$2731/2^{15}$
8	0.08333587646484	$10923/2^{17}$
9	0.08333396911621	$43691/2^{19}$

the equivalent fractions

$$\sigma^2_{0,i} = \frac{1 + 2\sum_{j=0}^{i-2} 2j}{2^{2i+1}} = \frac{1}{3}\frac{1 + 2^{2i-1}}{2^{2i+1}} = \frac{1}{12}\left(1 + \frac{1}{2^{2i-1}}\right) \quad (25)$$

Finally, the combination of (23) and (25) yields

$$\begin{aligned}\sigma^2_{f_{w_Q},\Delta f} &= \frac{1}{12}\left(1 + \frac{1}{2^{2\Delta f-1}}\right)2^{-2f_{w_Q}} \\ &= \frac{1}{12}\left(2^{-2f_{w_Q}} + 2^{-2f_{w_Q}-2\Delta f+1}\right) \\ &= \frac{1}{12}\left(2^{-2f_{w_Q}} + 2^{-2f_w+1}\right) \end{aligned} \quad (26)$$

which corresponds to the expression of the variance given in (13).